DOCUMENT RESUME

ED 346 115 TM 018 355

AUTHOR Huberty, Carl J.
TITLE Statistical Testing in the Behavioral Sciences: Textbook Developments.
PUB DATE Apr 92
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavioral Sciences; Content Analysis; *Educational Testing; Educational Trends; Evaluation Methods; Higher Education; Literature Reviews; Measurement Techniques; *Psychological Testing; Scaling; Social Science Research; *Statistical Analysis; *Test Use; *Textbook Content; Textbook Evaluation; Textbook Research
IDENTIFIERS P Values

ABSTRACT

Textbooks that have been and are being used by students in education and psychology to learn about statistical testing are reviewed. Twenty-eight textbooks published prior to 1950 were reviewed. These textbooks were found to focus on descriptive methods, to concentrate on educational testing, and to present limited formal statistical formulation. Five sets of textbook editions were also examined. Each set covered 20 to 40 years, with an average of about 30 years; and the study included editions appearing once every 7 years on the average. The sets are similar in that all used the p-value approach to statistical testing. Two sets retained the "old" structure that included a sizable discussion of validity, reliability, and scaling of performance tests, while the remaining three sets focused entirely on statistical methods. Seventeen recently published statistical methods textbooks were identified as possibly appropriate for behavioral and social science students. These textbooks are analyzed for the following: (1) logic used to introduce testing; (2) testing strategies; (3) test method used; and (4) inclusion of p-value. It is not statistical testing itself, but rather the presentation of statistical testing in textbooks and as taught that should be scrutinized. Four tables present lists of textbooks. There is a 34-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CARL J. HUBERTY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346115

Statistical Testing in the Behavioral Sciences: Textbook Developments

Carl J Huberty

University of Georgia

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992

TM018355

BEST COPY AVAILABLE

Introduction 1

Textbook Review 2

Pre-1950 Textbooks 2

Textbook Sets 7

Some Recent Textbooks 10

Comments 12

Levels of Significance 13

Importance of Effects 14

Statistical Power and Sample Size 15

Multiple Testing 15

Conclusion 16

References 17

Statistical Testing in the Behavioral Sciences: Textbook Developments

Statistical testing was applied by the English scholar John Arbuthnot nearly 300 years ago. [By no means should we credit the likes of R. A. Fisher, J. Neyman, E. S. Pearson, and K. Pearson with originating the concept of testing -- this, of course, in no way diminishes the contribution of these four, and other 20th century statisticians, toward developing and formalizing the concept.] Spanning about 200 years, from the early 1700s to the early 1900s, five applications of statistical testing were made in a variety of contexts; these are given in Table 1. Discussions of these situations are presented by Gigerenzer et al (1989), Hacking (1965), Porter (1986), and Stigler (1986). Statistical testing terms in current use today such as null hypothesis, critical value, and P value were not used in these applications, but a logic of testing that is consonant with that found in many statistical methods textbooks was evident.

An integral part of the logic used in these applications (and in subsequent logics expressed in textbooks) is that of probability. The view of probability employed during much of the past 300 years and by many people currently using and teaching statistical testing is that of relative frequency. As may be clear from readings on probability and testing, frequentists often disagree somewhat one from another. According to Savage (1961, p. 577), "The widest cleft between frequentists is that

Table 1
Statistical Testing Applications

Year	Person	Context
1710	Arbuthnot	Sex ratio at birth
1767	Michelle	Distribution of stars
1823	Laplace	Moon phase and barometric changes
1900	K. Pearson	Goodness of fit
1908	Gosset	A single mean

between R. A. Fisher and those who side closely with him on the one hand and those who more or less associate themselves with the school of Jerzy Neyman and Egon Pearson." The "more or less" phrase is particularly relevant to the teaching and learning of statistical testing, as will be indicated later in this paper.

The Fisher and Neyman-Pearson philosophies of statistical testing will not be discussed in detail here. The interested reader may refer to Oakes (1986, chap. 6), Salsburg (1990), Seidenfeld (1979, chaps. 2 and 3), and Spielman (1974) -- for a

response to Spielman, see Carlson (1976) -- for discussions and evaluations of the two schools of statistical inference. Very briefly, the Fisher significance testing approach involves a single (null) hypothesis with a strength-of-evidence statistic -- a P value or P(Data|H_0) value. The Neyman-Pearson hypothesis testing approach involves a null hypothesis (H_0) and alternative hypothesis (H_1), critical values(s), and a fixed value of the probability (alpha) of rejecting H_0 when H_0 is true. In a step-wise manner, the two approaches may be presented as follows:

Significance Testing	Hypothesis Testing
1. State H_0	1. State H_0 and H_1
2. Specify test statistic (T) and referent distribution	2. Specify test statistic (T) and referent distribution
3. Collect data and calculate value of T	3. Specify α value and determine rejection region (R)
4. Determine P value	4. Collect data and calculate value of T
5. Reject H_0 if P value is "small"; otherwise retain H_0	5. Reject H_0 in favor of H_1 if T value is in R; otherwise retain H_0

The former has been termed the P-value approach, while the latter has been termed the fixed-α approach. It is recognized that there may be a number of statisticians and methodologists who would contend that a contrast of the Fisher and Neyman-Pearson approaches to statistical testing is a nonissue. So be it.

The intent of this paper is to review some textbooks that have been and are being used by students in education and psychology for the purpose of learning about statistical testing. First, a number of textbooks published before 1950 will be briefly reviewed with respect to testing. Next, five sets of textbook editions will be discussed. Then, discussions of testing in some recent (post-1989) texts will be reviewed. The paper will conclude with some comments about testing issues and practices.

Textbook Review

Pre-1950 Textbooks

Twenty-eight textbooks that were published prior to 1950 were located and considered for review; they are listed in Table 2. The so-called P-value approach was described by R. A. Fisher

in his 1925 book (Statistical Methods for Research Workers) while the fixed-α approach was presented in the mid-1930s -- see the 1949 Johnson book (p. 68) for 1933 and 1936 Neyman-Pearson references. The decision to use 1950 as a cut-off is somewhat arbitrary, but the intent was to allow for time for the Neyman-Pearson approach to appear in textbooks. No claim is made that all textbooks that might have been used in the behavioral sciences were located; in fact, existence of additional textbooks is known but they were not in the UGA library. [Books published before 1950 within the five sets of textbook editions are discussed in the next section rather than in this section.]

The six pre-1920 textbooks listed in Table 2 have three characteristics in common: (1) a focus on descriptive methods; (2) an educational testing influence; and (3) limited formal statistical formulation. That the first and third characteristics coexist may be sensible. Each book presents a large number of examples, many of which come from the field of educational tests and measurement. It is common to find descriptive indices of location (mean, median, mode, and quartiles), of dispersion (standard deviation and probable

---

Table 2
Pre - 1950 Statistical Methods Textbooks

1910 Bowley, A. L. An elementary manual of statistics. London: MacDonald and Evans.
1910 Elderton, W. P. Primer of statistics. London: Black.
1913 Zizek, F. Statistical averages. New York: Holt.
1915 Bowley, A. L. An elementary manual of statistics (2 ed.). London: Macdonald and Evans.
1917 Rugg, H. O. Statistical methods applied to education. Boston: Houghton Mifflin.
1917 Secrist, H. An introduction to statistical methods. New York: Macmillan.

1921 King, W. I. The elements of statistical method. New York: Macmillan.
1923 Kelley, T. L. Statistical method. New York: Macmillan.
1924 Williams, J. H. Graphic methods in education. Boston: Houghton Mifflin.
1925 Odell, C. W. Educational statistics. New York: Century.
1925 Otis, A. S. Statistical method in educational measurement. New York: World.
1925 Secrist, H. An introduction to statistical methods (2nd ed.). New York: Macmillan.
1925 Thurstone, L. L. The fundamentals of statistics. New York: Macmillan.
1926 Bowley, A. L. Elements of statistics (5th ed.). London: King.

1928 Holzinger, K. S. Statistical methods for students in education. Boston: Ginn.
1929 Williams, J. H. Elementary statistics. New York: Heath.

1935 Odell, C. W. Statistical method in education. New York: Appleton-Century.
1936 Sorenson, H. Statistics for students of psychology and education. New York: McGraw-Hill.
1937 Bowley, A. L. Elements of statistics (6th ed.). London: Staples.
1937 Enlow, E. R. Statistics in education and psychology. New York: Prentice-Hall.
1938 Lindquist, E. F. A first course in statistics. Cambridge, MA: Houghton Mifflin.
1938 Wert, J. E. Educational statistics. New York: McGraw-Hill.
1939 Goulden, C. H. Methods of statistical analysis. New York: Wiley.
1939 Gray, C. T., & Votaw, D. F. Statistics applied to education and psychology. New York: Ronald.

1940 Lindquist, E. F. Statistical analysis in educational research. Boston: Houghton Mifflin.
1943 Walker, H. M. Elementary statistical methods. New York: Holt.
1947 Kelley, T. L. Fundamentals of statistics. Cambridge, MA: Harvard University Press.
1949 Johnson, P. O. Statistical methods in research. New York: McGraw-Hill.

---

error), and of association (expectancy tables in 1910 by Elderton, and (Karl) Pearson correlation in 1913 by Zizek). A "probable error" (P.E.) was defined as early as 1910 in the Elderton book. A probable error of a set of scores is defined as 0.6745 times the standard deviation of the set. Probable errors were often presented in the context of test score unreliability. [The constant 0.6745 is based on a consideration of the hypothetical standard normal score distribution; a range of (obtained score $\pm$ 1·P.E.) encompasses about 50% of the area under a normal curve.] The idea of a probable error of a score was extended to a probable error of a mean (see, e.g., the 1917 book by Rugg). Although the probable error of a mean was not at first explicitly associated with statistical testing, it may be considered as a forerunner for statistical inference. The P.E. of a mean was also at first presented in the context of discussions of unreliability; in effect what was given were 50% confidence intervals for the mean of a distribution.

Five of the listed books published in the 1920s -- the King, Otis, and Thurstone books plus the two books by Williams -- are

very much like the earlier books. The other five books of the 1920s listed in Table 2 contain discussions of probability and precision of estimators. Beyond basic ideas of probability, the probability distribution discussed is, as termed (in 1925 and 1917) by Secrist, the "normal law of error distribution." Precision is presented in terms of standard/probable errors of measures or of statistics. Hints at formal statistical testing are given by Odell (in 1925) and by Holzinger (in 1928) (fitting data to the normal frequency curve) and by Bowley (in 1926) and (in 1923) by Kelley (fitting data to Pearson curves). The earliest explicit reference to (Karl) Pearson's chi-squared goodness-of-fit test was given in 1928 by Holzinger. It was also in the Holzinger book that a "formal" interpretation of a P value was made -- "Unless the value of P be .2 or more, the fit cannot be regarded as good and other curves should be tried" (p. 247). [No reference to R. A. Fisher was found in the Holzinger book.] In comparing the difference of two obtained variable scores, the 1925 Odell book states that "the difference is probably, but not surely, significant" (p. 225).

Moving into the 1930s, there was a change of presentation of statistical testing in the textbooks. The 1938 Wert book was the only one still devoted to descriptive methods. As might be expected, none of the 1930s textbooks made any reference to the formal Neyman-Pearson approach to statistical testing. Statistical testing for three books (Odell in 1935, Bowley in 1937, and Enlow in 1937) still centered on the Pearson goodness-of-fit chi-squared test; the focus of interpretation was on tail area probabilities (i.e., P values). Although writings of R. A. Fisher were referred to by these two authors, the impact of Fisher's approach to statistical testing is not at all clear. There was no discussion of the goodness-of-fit test in the 1936 Sorenson book; he did, however, discuss the statistical significance of a difference between two means. Probable errors of mean difference, in relation to the standard normal probability distribution, were also considered in the Sorenson book. A similar discussion of mean difference significance was given in the 1939 Gray & Votow book.

Two of the eight books found and published in the 1930s deserve special mention. In the 1938 Lindquist book there was a chapter entitled, "Sampling Error Theory." This chapter centered on the "dependability or reliability" of parameter estimates. The notion of a "sampling distribution of the mean," along with the standard error of the mean, was described. In introducing the testing of a (null) hypothesis, Lindquist discusses the "degree of confidence" (p. 111) that is associated with a rejection of an hypothesis, and the tenability of an hypothesis -- what is used is, in essence, a P value. The hypothesis testing steps listed are clearly those of the (Fisher) P-value approach. Without referring to Neyman and Pearson, however, Lindquist discusses the use of a "critical value" (p. 132). Critical

values are discussed for normal-based tests, but also in the contexts of "small sample theory" -- this was the first book in which was found a use of the Student t family of probability distributions. An appendix table reported critical values, but no reference was made to "Student t."

The final 1930s book is that by Goulden (in 1939). This post-introductory textbook on statistical methods contains material that has a "direct bearing on the principles of the design of experiments." Goulden's background as an agricultural scientist might explain at least some of the topics covered: goodness of fit, analysis of variance, and analysis of covariance. Credit is given in the Preface to R. A. Fisher, who "has been mainly responsible for the development of the methods set forth" (p.vi). This is the first book listed in Table 2 in which reference to "Student" is given; according to Goulden, Student designated X/s by the letter z, and later designated X/s_x by the letter t. [It may of interest to some that Goulden presented ANOVA as a regression analysis (without the use of data coding schemes).]

Four books published in the 1940s, exclusive of those in the textbook sets discussed in the next section, were located. Presentations of statistical testing in these books will be reviewed in order of publication year. In the 1940s and 1950s the 1940 Lindguist textbook was considered an "advanced" book (about on the same level as the 1939 Goulden book). While no explicit reference to the Neyman-Pearson approach was found, Lindguist's steps for "testing a statistical hypothesis" (pp. 14-15) were clearly of that vein. [It is interesting that repeated references were made to the Fisher books, Statistical Methods for Research Workers and Design of Experiments.] This is the first book in which discussions of the "two types of error that are risked" (p.16) were given, though the expressions "Type I" and "Type II" and the Greek letters α and β were not used. Tests covered were: chi-squared tests of goodness of fit, independence, and homogeneity; Student t tests for a single mean, two independent means, and difference of related means; F test of a variance ratio, ANOVA, and ANCOVA; and tests for linear correlation and regression. The correlation ratio (i.e., eta) was presented in terms of ANOVA -- so, the idea of a correlational effect size has been around for 50 years!

Although Fisher's logic of testing is quoted in the 1943 Walker book (p. 263), the presented approach to testing is basically that of Neyman and Pearson. One- and two-sample tests are discussed. Pearson's goodness-of-fit test is not covered.

The 1947 Kelley book is quite different from all others published to that date. It is, as the author suggests, a treatise on the "philosophy and method" of statistics. There is no chapter or section of the book devoted to statistical testing

per se. Rather, various tests (e.g., Pearson's goodness-of-fit test, two-sample t test, ANOVA F test) are briefly mentioned throughout. In general, a P-value approach is advocated, but testing steps are not made explicit. There are many references to Fisher's writings; Fisher's influence in the Kelley book is fairly obvious. This is the first "applied" book in which relationships among the four families of probability distributions (normal, t, chi-squared, and F) are discussed in some detail.

Finally, the 1949 Johnson book was found to be the first textbook that might have been used by behavioral science students and instructors in which the fixed-α approach was followed "to the letter" in introducing statistical testing. Explicit reference was made to Neyman and Pearson (p. 64). The expression "critical region" as well as expressions "error of the first kind" and "error of the second kind" were used. "Introducing" was underscored above because Johnson often (e.g., on pp. 73, 82, 95) used--slipped in? -- the P-value approach in discussing results of testing. In particular, the goodness-of-fit test is discussed exclusively using P values (p. 93); that is, critical regions in terms of chi-squared probability distributions were not presented.

It appears that it took about 15 years -- from 1935 to 1950, roughly -- for the Neyman-Pearson philosophy to be integrated (to some extent) into presentations of statistical testing in behavioral science statistical methods textbooks. This is, perhaps, "par for the course."

Textbook Sets

The five sets of textbooks considered are alphabetically listed in Table 3. Familiarity with the sets was acquired by the current writer in four ways: (1) the 1959 edition of Ferguson was used in graduate study; (2) the 1978 edition of Guilford and the 1981 edition of Hays were used as textbooks in classes taught; (3) some editions of all five sets were used as occasional references for instruction; and (4) all editions in all sets except the 1971 Ferguson book and the 1926 Garrett book were scanned for purposes of the current paper.

The number of years covered by the textbook editions ranges from 20 (McNemar) to 40 (Garrett), with an "average" of about 30. The "average" number of years between editions is about 7. [With reference to the Garrett set, it is conjectured by the current writer that the appearance of six single-authored editions of an introductory-level textbook is history(!).]

The five sets of textbooks will now be briefly reviewed, in terms of statistical testing, in the order listed in Table 3.

Table 3
Five Sets of Behavioral Science Textbooks

Edition Years	
1959, 1966, 1971 1976, 1981, 1989	Ferguson, G.A. Statistical analysis in psychology and education. New York: McGraw-Hill.
1926, 1937, 1947 1953, 1958, 1966	Garrett, H. E. Statistics in psychology and education. New York: Longmans/McKay.
1942, 1950, 1956 1965, 1973, 1978	Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill.
1963, 1973, 1981 1988	Hays, W. L. Statistics (for psychologists). New York: Holt, Rinehart and Winston.
1949, 1955, 1962 1969	McNemar, Q. Psychological statistics. New York: Wiley

Note: Second authors were added in later editions of the Guilford and Ferguson books.

The introduction of testing by Ferguson in 1959 was clearly along the lines of Fisher in that the testing steps included one in which the null hypothesis is to be rejected if the observed result was "highly improbable." The four "logical steps" stated in the 1959 edition appeared nearly verbatim in the 1989 edition. A discussion of the two types of testing errors was first included in the 1966 edition and retained thereafter. The presentation of statistical testing was quite similar across the last five editions. Although critical values for test statistics are repeatedly alluded to in examples, no mention of critical region was found in any editions. A discussion of power of statistical tests was not given until the most recent (1989) edition (coauthored by Y. Takane).

The six editions by Garrett might be called the granddaddy of the five sets of textbooks listed in Table 3 -- that the books authored by one writer span 40 years is somewhat remarkable in and of itself! Garrett mentions (in the Preface) that new material in the 1947 edition included a chapter dealing with testing of experimental hypotheses. This is mentioned even though Pearson's goodness-of-fit chi-squared test was discussed in the 1937 edition and, I suspect, in the 1926 edition. [For whatever reason, textbook authors in the 1940s or so and earlier seemed not to consider the goodness-of-fit test to be a test of a statistical hypothesis, even though the concept of a null hypothesis appeared in textbooks in the 1930s (see, e.g., the 1938 Lindguist book listed in Table 2).] Garrett did, however,

make the connection between the goodness-of-fit test and null hypothesis explicitly in his 1953 edition. The 1947 edition may have started a textbook organization that appeared in many other textbooks in later years. This pertains to a discussion of the reliability (i.e., sampling variability and significance) of the difference between two means prior to a formal discussion of hypothesis testing -- see, for example, the Guilford set. Garrett's approach to statistical testing throughout the set -- at least for the last five editions -- was clearly the P-value approach. In fact, the latest edition (1966) did not even mention such concepts as statistical power, alternative hypothesis, and critical region. The 1966 edition, too, was still mentioning the use of a "critical ratio" -- a CR for large samples was a z statistic and a t statistic for small samples.

In the 1942 and 1950 editions by Guilford, analysis of variance is presented before "testing hypotheses." The discussion of testing in these first two editions is rather informal, and involved the use of tail areas. No mention of critical region or types of testing errors was made. The last four Guilford editions have chapters on significance of differences -- termed reliability of differences in the 1950 edition -- before a chapter on testing hypotheses. In the 1956 edition a recognition of two testing approaches is stated:

> Some investigators adopt a standard of significance in advance of the study or experiment. This lays down the rule for decision-making beforehand and makes it easy when the time comes. One disadvantage is that there may be temptation to modify the adopted standard after the results are in. Other investigators prefer not to adopt any rigid standard of acceptance or rejection of hypotheses. They are content to observe the level of significance reached and to report this fact. There is no need to follow either school of thought consistently (p. 215).

The usual Neyman-Pearson terms are presented; but test interpretations throughout the book are made via the use of P values. While discussing hypothesis testing in the sixth (1978) edition, Guilford and coauthor (B. Fruchter) use expressions like "the .01 level of confidence" and a "confidence level of .05" (p. 171) when referring to pre-set α levels. Three chapters in each of the last five Guilford editions were devoted strictly to measurement issues -- a carry-back to the statistical methods textbooks in the early 1900s?

The approach to statistical testing changed very little over the four editions by Hays. In general, a fixed α approach was followed, from introduction through later examples. No mention of P values was found in any of the four editions. All four editions have a section entitled, "Significance Tests and Common

Sense." In this section, statistical significance is defined in 1963 (p. 299) as a "statement about conditional probability" and in later editions as a "statement about the likelihood of the observed result." These statements could very well be interpreted as pertaining to P values.

McNemar's 1949 edition presented basically the Fisher approach to testing. As with many pre-1950 books there was not a text section on hypothesis testing. The most common test statistic was a "critical ratio" -- what current authors might call a z statistic -- with which is associated a "significance level represented by a P of .05" (p. 66). The two types of testing errors are defined (p. 69). A separate section on hypothesis testing was included in the 1955 edition. The testing approach used in this second edition was still basically a P-value approach; but now a critical region is defined, as is statistical power of a test (denoted by ß). With respect to testing, the 1962 and 1969 editions of McNemar were the same as the 1955 edition.

In sum, the five sets of textbooks are both similar and different. They are similar in that all sets, implicitly or explicitly, used the P-value approach to statistical testing. The only set that clearly favored the fixed-α approach -- but not to the complete disregard for the P-value approach -- is that by Hays. Also, at least four editions of each set were single-authored. Finally, for whatever it's worth, all five sets were published in New York City. They are different in that two sets -- Garrett and Guilford -- retained the "old" structure that included a sizable discussion of validity, reliability, and scaling of performance tests, while the remaining three sets focused entirely on statistical methods. The Hays set is the only set that deals exclusively with inferential methods; editions in the other four sets cover descriptive methods prior to inference. Content-wise, the Hays set might be considered an "outlier."

Some Recent Textbooks

Seventeen recently published statistical methods textbooks were identified that may be considered appropriate for behavioral and social science students -- see Table 4. These books are reviewed in terms of their discussions of statistical testing. The reviewing process centered on the testing approach advocated by the author(s). That is, was the P-value approach or the fixed-α approach suggested? Or, was some sort of hybrid of the two approaches suggested? [The intent was not to evaluate the quality of the textbooks, but merely to describe what was found.]

Table 4
Some Recent Textbooks

1990 Arney, W. R. Understanding statistics in the sciences. New York: Freeman.
1990 Freedman, D. et al. Statistics (2nd ed.). New York: Norton.
1990 Groninger, L. D. Beginning statistics within a research context. New York: Harper & Row.
1990 Hamilton, L. C. Modern data analysis. Pacific Grove, CA: Brooks/Cole.
1990 Jaccard, J., & Becker, M. A. Statistics for the behavioral sciences (2nd ed.). Belmont, CA: Wadsworth.
1990 May, R. B. et al. Application of statistics in behavioral research. New York: Harper & Row.
1990 McCall, R. B. Fundamental statistics for behavioral sciences (5th ed.). New York: Harcourt Brace Jovanovich.
1990 Rosenberg, K. M. Statistics for behavioral sciences. Dubuque, IA: Brown.
1990 Sprinthall, R. C. Basic statistical analysis (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
1990 Wampold, B. E., & Drew, C. J. Theory and application of statistics. New York: McGraw-Hill.
1990 Weinberg, S., & Goldberg, K. Statistics for the behavioral sciences. New York: Cambridge.

1991 Christensen, L. B., & Stoup, C. M. Introduction to statistics for the social and behavioral sciences (2nd ed.). Pacific Grove, CA: Brooks/Cole.
1991 Lehman, R. S. Statistics and research design in the behavioral sciences. Belmont, CA: Wadsworth.
1991 Welkowitz, J. et al. Introductory statistics for the behavioral sciences (4th ed.). New York: Harcourt Brace Jovanovich.

1992 Gravetter, F. J., & Wallnau, L. B. Statistics for the behavioral sciences (3rd ed.). New York: West.
1992 Heiman, G. W. Basic statistics for the behavioral sciences. Boston: Houghton Mifflin.
1992 Popham, W. J., & Sirotnik, K. A. Understanding statistics in education. Itasca, IL: Peacock.

Thus, four items were considered:
1. Logic used to introduce the concept of testing;
2. Test strategy given in formal testing steps (if listed);
3. Test method used in subsequent examples; and
4. Inclusion of "P value" in textbook index.

Testing logic used. Fourteen of the 17 books employ the idea of a probability "tail area" to introduce the concept of

testing. The Sprinthall book, the Christensen & Stoup book, and the Gravetter & Wallnau book are the exceptions; these three textbooks delved straight into the fixed-α approach. The predominant idea used was that of discarding a conjecture (or, hypothesis) if an extremely rare, or chance, or highly unlikely sample outcome was obtained.

Testing steps. Explicit lists of steps for testing were not found in five textbooks -- May et al, Sprinthall, Christensen & Stoup, Welkowitz et al, and Popham & Sirotnik. Only two of the 17 textbooks (Freedman et al and Hamilton) included the determination of a P value as one of the steps. Two textbooks -- Arney and Rosenberg -- included a step of comparing a probability tail area with a set α value; neither "P" nor "p" was used to denote the tail area in either textbook. The remaining eight textbooks presented steps that clearly reflected the fixed-α approach -- this is the approach, I believe, that dominated the textbooks in the 1960s, 1970s, and 1980s.

Test method used. In discussions of the two-independent-groups test for a mean difference, all but three textbooks used the fixed-α approach to testing. Consistent with earlier discussions, the Freedman et al and Hamilton textbooks presented the P-value approach and the Arney textbook used $P<\alpha$ in the two-independent-groups context. So, even though many textbooks introduce the concept of testing via tail areas, many of the same textbooks present the fixed-α approach in many testing contexts.

Inclusion of "P value". Only three textbooks included "P value" or "p value" in their indexes: Freedman et al, Hamilton, and Wampold & Drew. It may be surprising that even though Arney clearly used the idea of a tail area in statistical testing, the usual symbol (P or p) was not used. The limited discussion of P value in behavioral science textbooks may also be surprising in light of the common (conventional?) practice of referring to P values in journal articles. Incidently, in this day and age when computer software is so widely accessible, why inequalities like "P<.05" are used in journal articles in preference to, say, "P=.029" is somewhat puzzling.

Comments

Since the 1960s, the roles and practices of statistical testing in quantitative research have been discussed in a number of substantive disciplines; for example, in economics (Good, 1981; Mayo, 1981), education (Carver, 1978; Orey, Garrison, & Burton, 1989), marketing (Sawyer & Peter, 1983), medicine (Browner & Newman, 1987), occupational therapy (Ottenbacher, 1985), psychology (Meehl, 1975; Rosnow & Rosenthal, 1989; Serlin, 1987), social research (Guttman, 1985), and sociology (Gold, 1969). Testing is also currently a topic of interest to methodologists; recent articles have appeared dealing with new ideas (e.g., Simon & Bruce, 1991, on using bootstrap methods to

validate test results), with the rehashing of old ideas (e.g., Pillemer, 1991), and with the teaching of statistical testing (e.g., Bangdiwala, 1989). In the recent and excellent reference book for sociobehavioral sciences, Pedhazur and Schmelkin (1991, pp. 197-210) give about 60 references to writings related to the logic of statistical testing, P values, substantive importance, and a decision-based strategy that stems from, but is more extensive than, the Neyman-Pearson approach. The use of statistical tests has also been a concern of statisticians; a portion of an issue of a journal, Communications in Statistics - Theory and Methods, was devoted to the topic -- the first of four articles was written by nonother than Jerzy Neyman (1976).

There are various issues and concerns related to the practice and teaching of statistical testing. Among these are: (1) levels of significance; (2) importance of effects; (3) statistical power and sample size; and (4) multiple testing. Interpretations and perceptions of these are briefly discussed, in order.

Levels of Significance

Many textbooks authored in the 1990s suggest, by their examples, choosing a significance level (i.e, α) prior to data collection. This includes advanced books as well as introductory books. The chosen level -- usually α=.05 -- is used for a number of test statistics and a variety of design situations. That is, authors often stick by the .05 level regardless of test sensitivity, number of observations/units, number of variables, variable measures, and the number of tests involved. Sometimes -- very often? -- authors cite writings of R. A. Fisher as support for the choice of α=.05. Yet, Fisher (1959, p. 42) himself stated that, "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." As might be gathered from earlier statements in this paper, the current author does not subscribe to the notion of the fixed-α approach to testing. If a critical value, using α=.05, is 2.53 and a test-statistic value of 2.60 is obtained, a fixed-α researcher would conclude that a statistically significant result occurred. What if a test-statistic value of 2.45 is obtained? How about 2.40? Etc. If a value of 9.38 were obtained, would one still conclude statistical significance at the .05 level?

The role of P values in statistical testing has gone from one of major emphasis, as indicated in textbooks in the 1940s and earlier, to very limited emphasis from the 1950s through the 1970s, and then back to at least some emphasis in the 1980s. Do the 17 1990s textbooks listed in Table 4 reflect the current emphasis in teaching statistical testing, that is, little emphasis on P values? There was one textbook in the mid-1980s (Mosteller, Fienberg, & Rourke, 1983) and one in the late 1980s

(Moore & McCabe, 1989) that put considerable emphasis on P values. Two recent textbooks that also emphasize the P-value approach are the 1990 Freedman et al and Hamilton books listed in Table 4. Moore and McCabe (1989, p. 512) go so far as to imply that tables of critical values of probability distributions ubiquitously found in current textbooks will soon be unnecessary, because of the widespread availability of computer software to determine P values.

Importance of Effects

Of course, empirical researchers should not rely exclusively on statistical significance to assess results of statistical tests. Some type of measurement of magnitude or importance of the effects should also be made. This was recognized as far back as the 1930s in textbooks -- see the 1938 Lindguist book (p. 132) and the 1936 Sorenson book (p. 335) listed in Table 2. This notion was also expressed for the statistics community (e.g., Hodges & Lehmann, 1954) and for the lay reader (Moroney, 1951, pp. 218-219). In discussing statistical tests, Moroney states that, "Practical considerations must always be paramount. We must never lose sight of common sense and all those other relevant factors which cannot possibly be taken care of statistically "(p. 218). The use of common sense and sound judgment in the interpretation of data analysis results is certainly essential, but it would be helpful to have some kind of quantitative index in assessing the importance of effects. In recent decades such a type of index (omega) was suggested to behavioral science researchers in the 1963 edition of the Hays set listed in Table 3. In a group comparison setting the omega index yields an estimate of the relationship between the grouping variable and the outcome variable. The idea of a relationship index as an index of importance of effects in a group-comparison context was suggested over 50 years ago in the 1940 book by Lindguist (see Table 2) and in the 1950 edition of the Guilford set (see Table 3) -- both of these authors used eta instead of omega.

The quantitative assessment of importance of effects was popularized in the 1970s when the term "effect size" (ES) was bandied about the behavioral sciences. The interpretation usually given to ES was a standardized difference. A number of indices for assessing effect importance are reviewed by Maxwell and Delaney (1990, pp. 97-107). The current author judges that such an assessment should be routinely reported. This judgment is not, however, shared by some current textbook authors. Of the 17 books listed in Table 4, only seven (Groninger, Jaccard & Becker, May et al, Wampold & Drew, Weinberg & Goldberg, Welkowitz et al, and Heiman) even hint at such an index for group comparison studies; only the May et al book suggests the use of such an index in group contrast analyses.

Statistical Power and Sample Size

It is generally thought that Fisher did not consider the two types of statistical testing errors in his testing approach. This may not be too surprising because it is also generally thought that Fisher's testing ideas focused only on the null hypothesis. As such, then, the notion of statistical power did not, on the surface at least, seem to be a concern of Fisher. It has been argued that the Neyman-Pearson approach is preferable to the Fisher approach because the former takes statistical power of a test into explicit consideration. This explicit consideration would surely seem to be real as one reads the many textbooks written in the past four decades. Reason for the consideration of statistical power taken in many textbooks is, however, not entirely clear. Sure, textbooks discuss probabilities of committing a Type II error and the meaning and, sometimes, the calculation of its complement, statistical power. [Why some textbooks discuss statistical power as though there is only one power value for a test is not clear.]

Nearly all of the books listed in Table 4 discuss the effect sample size may have on the statistical power of a test. A few books present a formula for sample size, usually only for the one-sample case. The 1991 Welkowitz et al book presents formulas for a number of tests. The 1990 Jaccard & Becker book is an exception. This book contains sample-size tables for seven different statistical test situations. Using input of alpha, power, and eta-squared values, the reader may determine an approximate desired sample size.

Merely discussing the effect sample size may have on power is not enough. It is conjectured that subsequent to such textbook discussions, the student will forget about power and go through with his/her statistical test. A design-and-testing strategy that takes into consideration most if not all relevant aspects of the process and which is sort of a hybrid of the Fisher and Neyman-Pearson approaches has been suggested by Huberty (1987). [The association of Fisher and Neyman-Pearson with significance testing and hypothesis testing, respectively, was inadvertently reversed in the abstract of that article.]

Multiple Testing

The consideration of results from more than one statistical test may take place in three somewhat different contexts. One context is when multiple tests are conducted on a given data set with a single outcome variable. Two examples of this are the three omnibus tests in a two-factor analysis of variance, and the six pairwise group contrast tests for a one-way design with four groups. A second context is when a number of bivariate correlation coefficients are tested for significance. For example, for a five-variable problem, 10 tests would be conducted. A third context in which multiple tests might be considered is when results from a number of different studies are

to be combined. This problem has surfaced within the past 20 years in meta-analysis studies, but the problem has been discussed in the behavioral sciences for over 50 years -- see, for example, the 1940 Lindguist book (pp. 46-47) listed in Table 2.

No matter what multiple testing context is present, the fact that multiple tests are conducted should not be ignored. That is, each test among a number of tests should not be assessed as though it was the only test conducted. Rather, some modification or adjustment in the testing process should be seriously considered. A popular modification of late has been some type of Bonferroni adjustment. Only three of the books listed in Table 4 (May et al, Wampold & Drew, and Lehman) even mention any type of Bonferroni method; in all three books the discussion is restricted to multiple group contrast tests. A discussion of a number of Bonferroni adjustment methods that may be used in any context is given by Li, Huberty, and Olejnik (1992).

Conclusion

It is not statistical testing itself, but rather the textbook presentations, the teaching, the applications of testing, and the reporting of test results that should be scrutinized.

References

Bangdiwala, S. I. (1989). The teaching of the concepts of statistical tests of hypotheses to non-statisticians. Journal of Applied Statistics, 16, 355-361.

Browner, W. S., & Newman, T. B. (1987). Are all significant P values created equal? Journal of the American Medical Association, 257, 2459-2463.

Carlson, R. (1976). Discussion: The logic of tests of significance. Philosophy of Science, 43, 116-128.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Fisher, R. A. (1959). Statistical methods and scientific research. (2nd ed.). New York: Hafner.

Gigerenzer, G. et al (1989). The empire of chance. New York: Cambridge.

Gold, D. (1969). Statistical tests and substantive significance. The American Sociologist, 4, 42-46.

Good, I. J. (1981). Some logic and history of hypothesis testing. In J. C. Pitt (Ed.), Philosophy in economics (pp. 149-174). Boston: Reidel.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. Applied Stochastic Models and Data Analysis, 1, 3-10.

Hacking, I. (1968). Logic of statistical inference. New York: Cambridge.

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. Journal of the Royal Statistical Society, Series B, 16, 261-268.

Huberty, C. J (1987). On statistical testing. Educational Researcher, 16(8), 4-9.

Li, J., Huberty, C. J, & Olejnik, S. F. (1992, April). Multiple testing with modified Bonferroni methods. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data. Belmont, CA: Wadsworth.

Mayo, D. (1981). Testing statistical testing. In J. C. Pitt (Ed.), Philosophy in economics (pp. 175-203). Boston: Reidel.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Moore, D. S., & McCabe, G. P. (1989). Introduction to the practice of statistics. New York: Freeman.

Moroney, M. J. (1951). Facts from figures. New York: Penguin.

Mosteller, F., Fienberg, S. E., & Rourke, R. E. K. (1983). Beginning statistics with data analysis. Reading, MA: Addison-Wesley.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. Communications in Statistics - Theory and Methods, A5, 737-751.

Oakes, M. (1986). Statistical inference. New York: Wiley.

Ottenbacher, K. (1985). The significance of power and the power of significance: Recommendations for occupational research. Occupational Therapy Journal of Research, 4, 37-49.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis. Hillsdale, NJ: Erlbaum.

Pillemer, D. B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. Educational Researcher, 20(9), 13-17.

Porter, T. M. (1986). The rise of statistical thinking 1820-1900. Princeton, NJ: Princeton University Press.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Salsburg, D. (1990). Hypothesis versus significance testing for controlled clinical trials: A dialogue. Statistics in Medicine, 9, 201-211.

Savage, L. J. (1961). The foundations of statistics reconsidered. In J. Neyman (Ed.), Proceedings of the fourth symposium on mathematical statistics and probability, Vol. 1 (pp. 575-586). Berkeley, CA: University of California.

Sawyer, A. G., & Peter, J. P. (1983). The significance of statistical significance tests in marketing research. Journal of Marketing Research, 20, 122-133.

Seidenfeld, T. (1979). Philosophical problems of statistical inference. Boston: Reidel.

Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. Journal of Counseling Psychology, 34, 365-371.

Simon, J. L., & Bruce, P. Resampling: A tool for everyday statistical work. Chance, 4, 22-32.

Spielman, S. (1974). The logic of tests of significance. Philosophy of Science, 41, 211-226.

Stigler, S. M. (1986). The history of statistics. Cambridge, MA: Belknap.